Vorlesung 11b

Statistik von Anteilen

Ein Beispiel:

Große Population (♀ und ♂)

mit unbekanntem Weibchenanteil p

In einer Stichprobe vom Umfang $n = 53$ waren

23 Weibchen und 30 Männchen

Wie kann man p “mit Konfidenz” schätzen?

Hat man Grund, an der Hypothese “$p = 1/2$” zu zweifeln?

Modellvorstellung:

rein zufälliges Ziehen

mit Zurücklegen:

Münzwurf

$X = (X_1, \ldots, X_n)$ mit unbekanntem p

$k/n = 23/53$ wird gedeutet

als Ausgang der Zufallsvariablen

$$\widehat{p} := \frac{1}{n}(X_1 + \cdots + X_n) \ .$$

Wie groß p auch immer ist:

Der *Schätzer* $\widehat{p}$ hat Erwartungswert p

und Standardabweichung $\sigma/\sqrt{n}$, mit

$$\sigma = \sigma(p) := \sqrt{p(1-p)} \, .$$

Schätzung für σ:

$$\widehat{\sigma} := \sqrt{\widehat{p}(1-\widehat{p})} \, .$$

Im Beispiel ergibt sich für $\widehat{p} \pm \widehat{\sigma}/\sqrt{n}$

der Wert 0.43 ± 0.068

typische Schwankung in der Schätzung von p

Es geht noch besser:

De Moivre - Laplace sagt:

Für nicht all zu kleines $np(1-p)$ ist

$$\widehat{p} \sim \; p + \frac{\sigma}{\sqrt{n}} Z$$

mit Z $\quad N(0,1)$ - verteilt

$$\frac{\widehat{p}-p}{\sigma/\sqrt{n}} \sim Z$$

also auch

$$\frac{\widehat{p}-p}{\widehat{\sigma}/\sqrt{n}} \sim Z$$

$$\mathbf{P}_p\Big(-2 \le \frac{\widehat{p}-p}{\widehat{\sigma}/\sqrt{n}} \le 2\Big) \approx \mathbf{P}(-1.96 \le Z \le 1.96) = 0.95\,.$$

$$\Big\{-2 \le \frac{\widehat{p}-p}{\widehat{\sigma}/\sqrt{n}} \le 2\Big\} = \Big\{\widehat{p} - 2\frac{\widehat{\sigma}}{\sqrt{n}} \le p \le \widehat{p} + 2\frac{\widehat{\sigma}}{\sqrt{n}}\Big\}$$

$$I := \Big[\widehat{p} - 2\frac{\widehat{\sigma}}{\sqrt{n}}, \widehat{p} + 2\frac{\widehat{\sigma}}{\sqrt{n}}\Big]$$

$$\mathbf{P}_p(p \in I) \approx 0.95$$

Man nennt I ein *Konfidenzintervall* für p,

approximativ zum *Niveau* 0.95

Zum Merken:

Das Ereignis

$$\left\{|\widehat{p}-p| \leq 2\widehat{\sigma}/\sqrt{n}\right\}$$

hat ungefähr W'keit 0.95

(weil $\widehat{p}$ approximativ normalverteilt ist

mit EW p und Standardabweichung $\sigma/\sqrt{n}$)

Also überdeckt das zufällige Intervall

$$I = \left[\widehat{p} - 2\frac{\widehat{\sigma}}{\sqrt{n}}, \widehat{p} + 2\frac{\widehat{\sigma}}{\sqrt{n}}\right]$$

den "wahren Parameter" p mit W'kt ungefähr 0.95.

In unserem Beispiel ergab sich das Intervall $[0.29, 0.57]$.

Verträgt sich die Hypothese “$p = 1/2$“ mit den Daten?

Die beobachtete Abweichung war

$$|k/n - 1/2| = |23/53 - 1/2| \approx 0.07$$

Wie wahrscheinlich ist unter unserer Hypothese eine

(mindestens) so große Abweichung ?

Für $p = 1/2$ ist $\sigma_{\widehat{p}} = 1/\sqrt{4n}$,

in unserem Beispiel ist das 0.069.

$$\mathbf{P}(|Z| > 1) = 0.32$$

So etwas wird jedes dritte Mal vorkommen ...

Maximum-Likelihood-Schätzer am Beispiel des Münzwurfs:

$(X_1, \ldots, X_n)$ sei p-Münzwurf mit unbekanntem p

Beobachtet wird die Realisierung $(a_1, \ldots, a_n)$

mit $k = a_1 + \ldots + a_n$.

Unter allen p ist k/n derjenige Parameter, mit dem

$$\mathbf{P}_p(X_1 = a_1, \ldots, X_n = a_n) = p^k(1-p)^{n-k}$$

maximal ist.

Man sagt:

$$\frac{1}{n}(X_1 + \cdots + X_n)$$

ist der Maximum-Likelihood-Schätzer für p.

Für $k = n$ (kein Misserfolg in n Versuchen) ergibt sich 1 als Maximum-Likelikood-Schätzung von p.

Das ist möglicherweise zu optimistisch.

Eine Alternative bietet der sogenannte *Bayes-Schätzer* (vgl Buch S. 127).

Hier denkt man an ein zweistufiges Experiment:

1. eine auf $[0, 1]$ uniform verteilte Zufallsvariable U
2. gegeben $\{U = u\}$ einen Münzwurf mit Erfolgswahrscheinlichkeit U.

$$\tilde{p} := \mathbf{E}[U|K] \ .$$

Erinnerung an Vorlesung 9b:

$Z_1, Z_2, \ldots$ sei ein Münzwurf mit uniform auf $[0,1]$ verteiltem zufälligem Erfolgsparameter U,

K_n sei die Anzahl der Erfolge in den ersten n Versuchen.

Gefragt ist nach $\mathbf{E}[U \mid K_n = k]$.

Wir wissen schon:

Die bedingte Dichte von U gegeben $\{K_n = k\}$ ist

$$\mathbf{P}_k(U \in du) = \frac{1}{\frac{1}{n+1}} \binom{n}{k} u^k (1-u)^{n-k} \, du$$

$$\mathbf{E}[U \mid K_n = k] = \int \frac{1}{\frac{1}{n+1}} \binom{n}{k} u^{k+1} (1-u)^{n-k} \, du$$

$$= \frac{k+1}{n+2} \int \frac{1}{\frac{1}{n+2}} \binom{n+1}{k+1} u^{k+1} (1-u)^{(n+1)-(k+1)} \, du$$

$$= \frac{k+1}{n+2}.$$

Man nennt dies auch den

Bayes-Schätzer für die Erfolgswahrscheinlichkeit

(bei a priori uniform verteilter Erfolgswahrscheinlichkeit).

Eine Beziehhung zur Pólya-Urne:

Im Buch S. 113/114 liest man nach:

Ein Münzwurf $(Z_1, Z_2, \ldots)$

mit uniform verteilter Erfolgswahrscheinlichkeit U

ist so verteilt wie die Folge der Zuwächse in Richtung Osten

in einer Nordost-Wanderung à la Pólya.

Also:

$$\mathbf{E}[U|K_n] = \mathbf{P}[Z_{n+1} = 1|K_n] = \frac{k+1}{n+2}.$$